

Genotype to phenotype

09.01.14

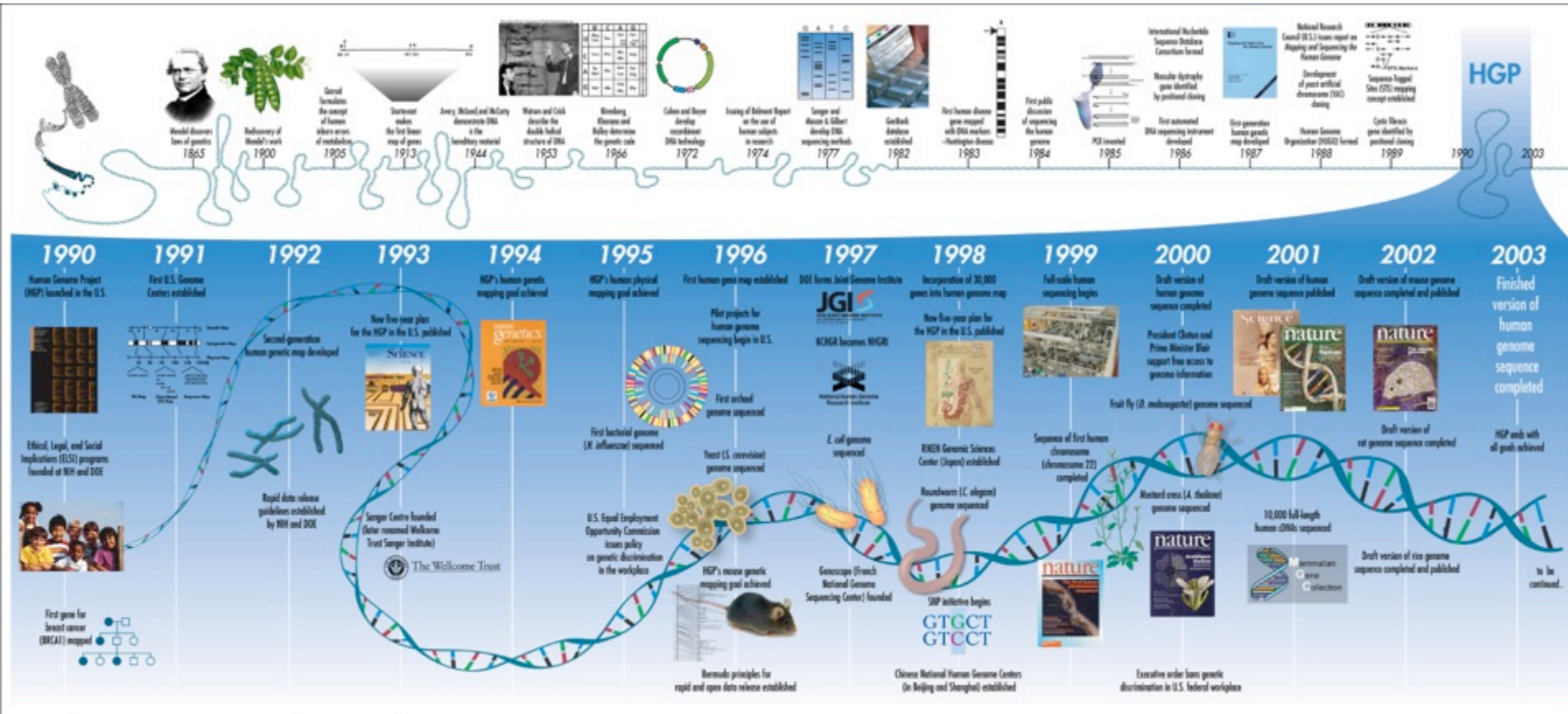
Valborg Gudmundsdottir

Outline for the afternoon

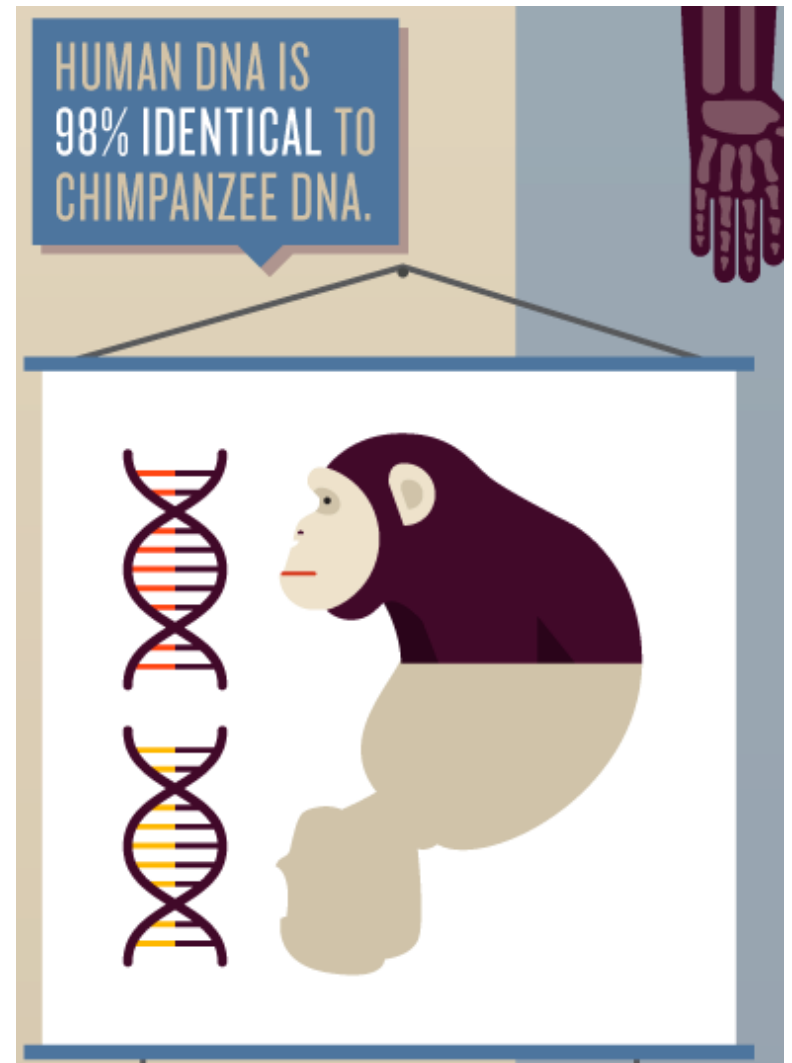
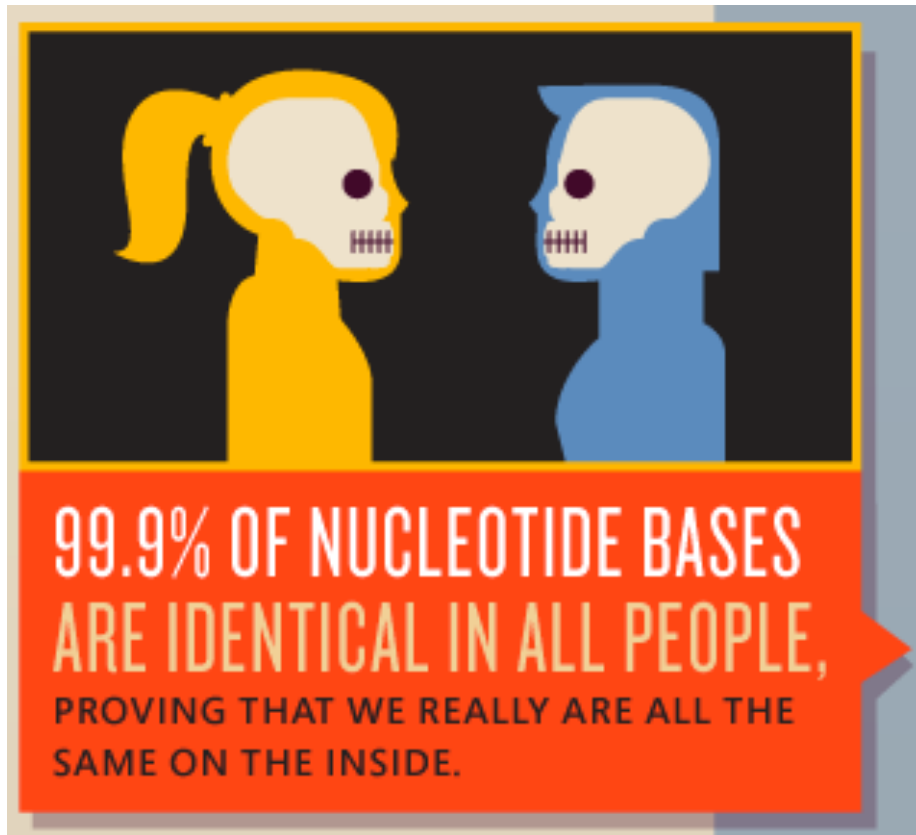
- The human genome
- Genomic variation
- Genotypes, homozygotes and heterozygotes
 - *Short exercise*
- Genotyping
- Phenotypes
- Associating genotype with phenotype, genome-wide associated studies (GWAS)
- Genotype to phenotype predictions
 - *Main exercise*

THE HUMAN GENOME

The Human Genome Project




The Human Genome Project



The Human Genome Project



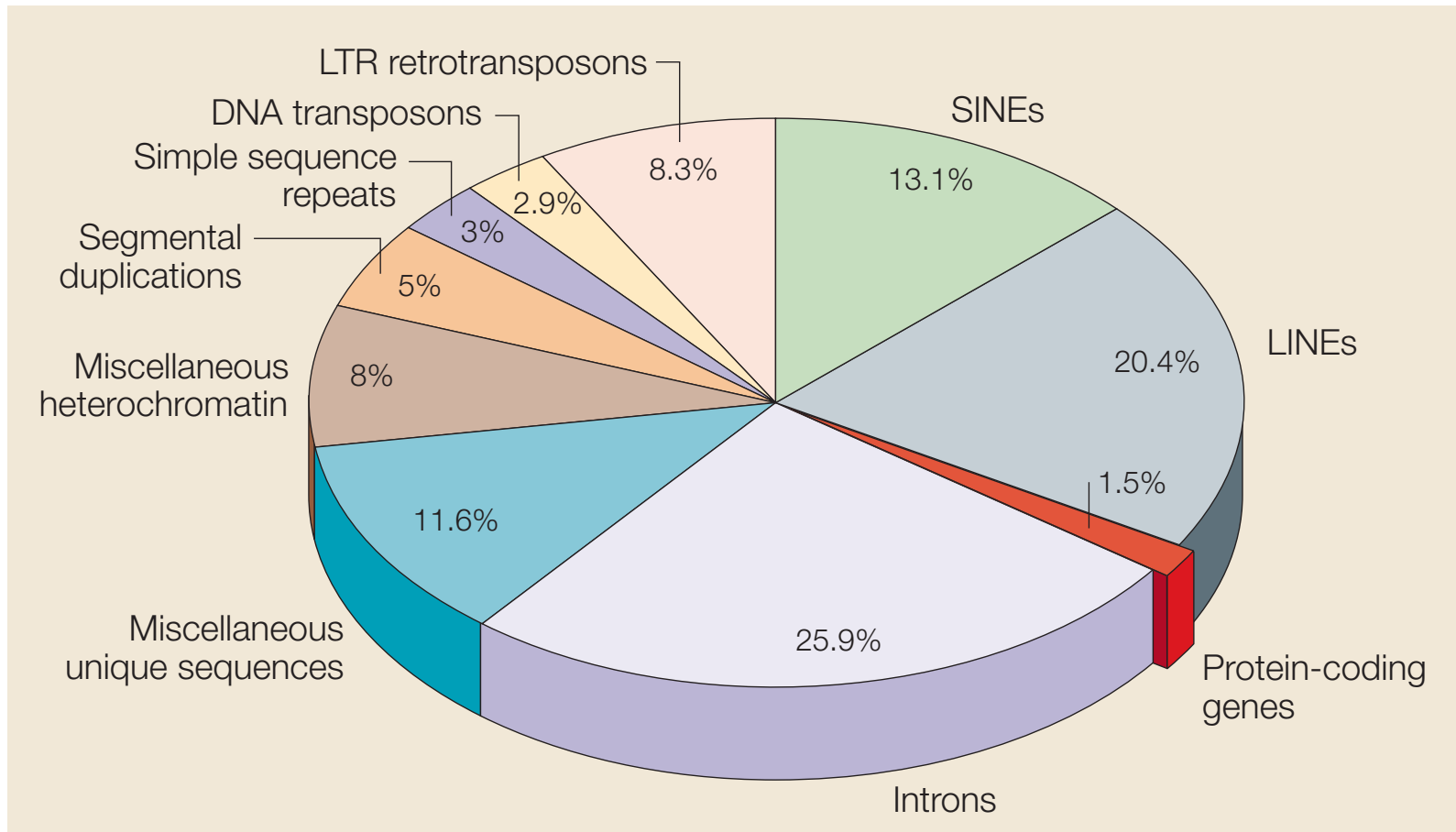
THE HGP IDENTIFIED THE APPROXIMATELY
25,000 GENES IN HUMAN DNA.



THE VAST MAJORITY OF DNA IN THE HUMAN GENOME,
**(97%) CONSISTS OF NON-GENETIC
SEQUENCE WITH UNKNOWN FUNCTION,**
OFTEN CALLED "JUNK DNA."

The illustration depicts a dark room with a blue door on the left. Inside, there are several cardboard boxes, some open, and a yellow box with a red circle. A yellow box with a red circle is also visible. The scene is cluttered, representing the 'junk DNA' mentioned in the text.

Main components of the eukaryotic genome



GENOMIC VARIATION

Genomic variation

- Single nucleotide polymorphisms (SNPs)
- Insertions/deletions
- Copy number variations (large)
- Variable (short) number tandem repeats

Single nucleotide polymorphisms (SNPs)

A single nucleotide (A,T,C,G) DNA sequence alteration

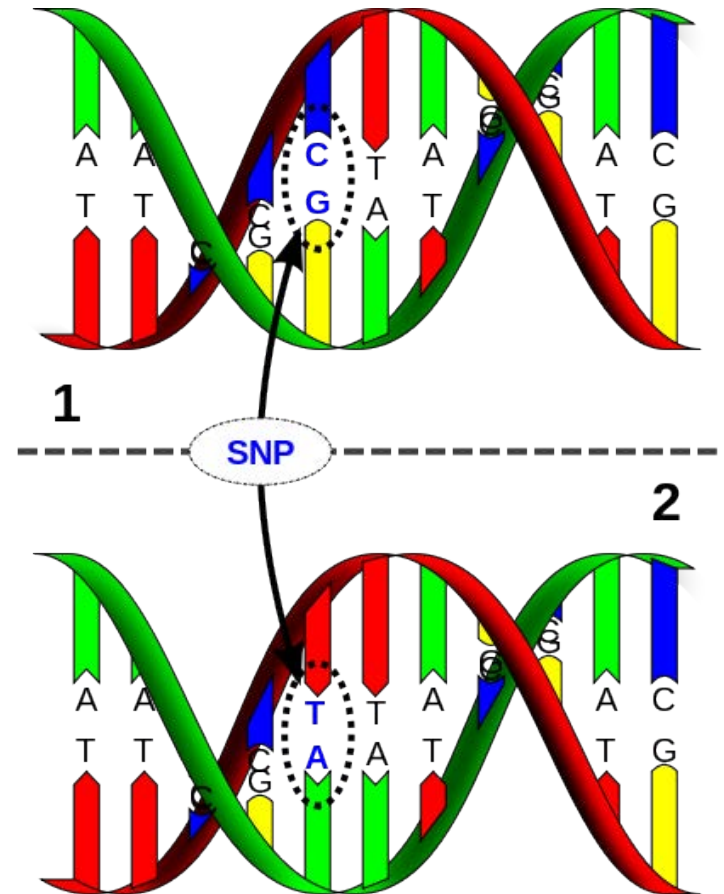
... **A**CGGCTAA ...

... **A****T**GGCTAA ...

C and T are the **alleles**
for this position

DNA is double stranded

- “C” or “T” on red strand
- “G” or “A” on green strand

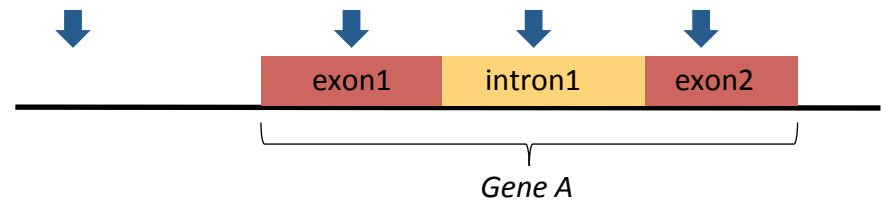


Single nucleotide polymorphisms (SNPs)

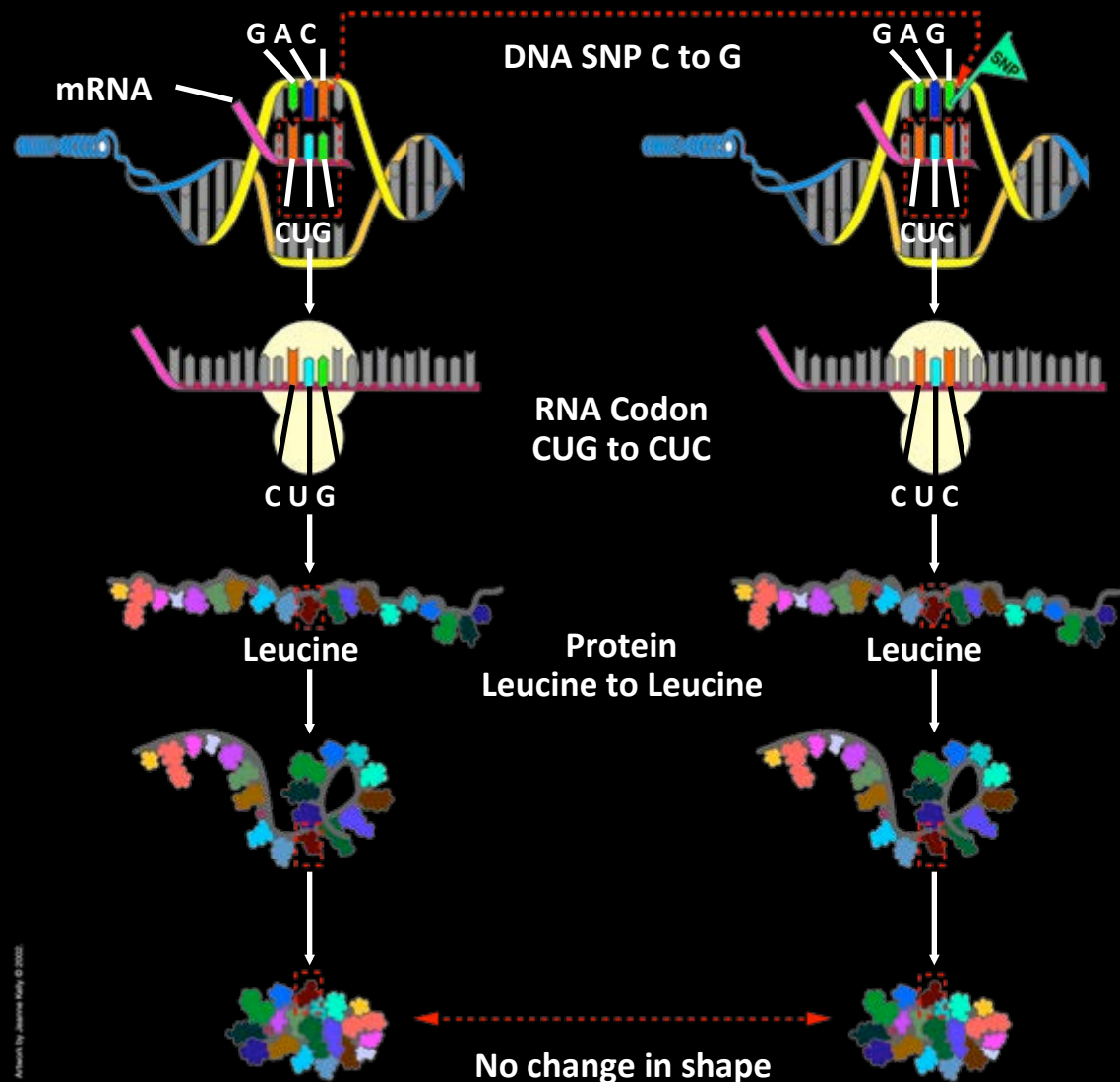
- Occur in at least 1% of the population
- Most common kind of human genetic variation
- 10-30 million SNPs in the human genome
- Occur every 100-300 bases along the 3-billion-base human genome
- Evolutionary stable

Single nucleotide polymorphisms (SNPs)

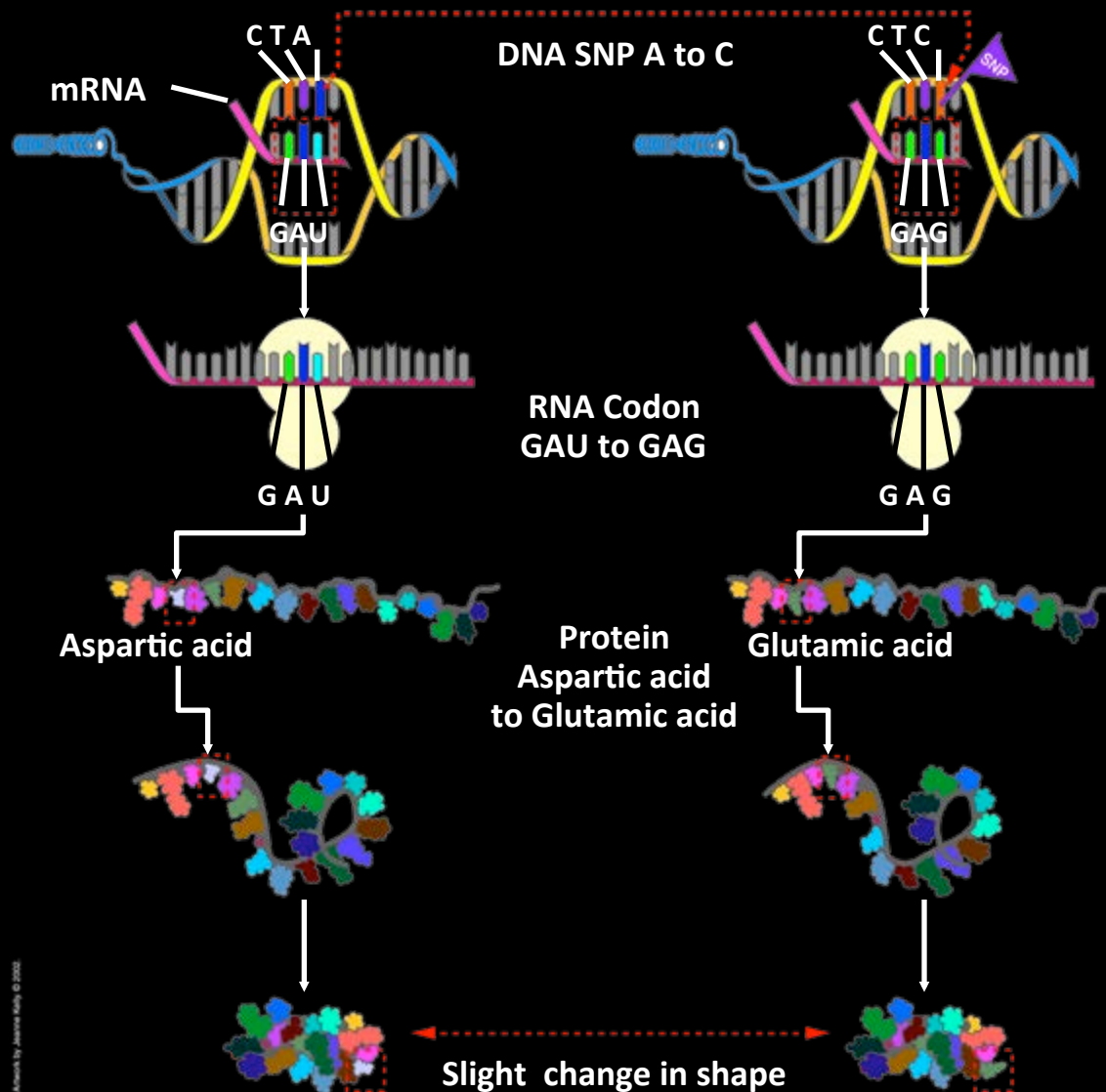
- Non-coding region
- Coding region
 - Synonymous
 - Nonsynonymous



SNPs in Coding Regions – Synonymous: No Changes in Protein



SNPs in Coding Regions – Nonsynonymous: Changes in Protein



Adapted by Jennifer Kelly, © 2002

dbSNP database

- rs numbers
- chromosome and positions
- Strand orientation

Reference SNP(refSNP) Cluster Report: rs17822931 ** With probable-pathog... "http://www.ncbi.nlm.nih.gov/sites/varvu?gene=85320&rs=17822931"> [detail] **

http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=17822931

NCBI dbSNP Short Genetic Variations

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP : for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive! Go

GENERAL

RSS Feed

Contact Us

Site Map

dbSNP Homepage

Announcements

dbSNP Summary

FTP Download

HUMAN VARIATION

SNP SUBMISSION

DOCUMENTATION

SEARCH

RELATED SITES

Reference SNP(refSNP) Cluster Report: rs17822931 ** With probable-pathogenic allele [detail] **

RefSNP

Organism: human ([Homo sapiens](#))

Molecule Type: Genomic

Created/Updated in build: 123/135

Map to Genome Build: [37.3](#)

Validation Status:

Citation: [PubMed](#)

Allele

Variation Class: SNV: single nucleotide variation

RefSNP Alleles: C/T

Allele Origin: G: Germline
A: Germline

Ancestral Allele: C

Clinical Source:

Clinical Significance: **With probable-pathogenic allele** [\[detail\]](#)

MAF/MinorAlleleCount: T=0.310/679

MAF Source: 1000 Genomes

HGVS Names

NC_000016.9:g.48258198C>T

NG_011522.1:g.15891G>A

NM_032583.3:c.538G>A

NM_033151.3:c.538G>A

NM_145186.2:c.538G>A

NP_115972.2:p.Gly180Arg

NP_149163.2:p.Gly180Arg

NP_660187.1:p.Gly180Arg

Links, Linkout


SNP Details are organized in the following sections:


[GeneView](#) [Map](#) [Submission](#) [Fasta](#) [Resource](#) [Diversity](#) [Validation](#)

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer) ↑

Assembly	Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh37.p5	37.3	16	48258198	NT_010498.15	1872387	+	C	+	view	blast
reference	36.3	16	46815809	NT_010498.15	1872387	+	C	+	view	blast
Celera	36.3	16	32765323	NW_926462.1	1830122	+	C	+	view	blast
hg19	37.2	16	34149352	NM_001028288.2	3012285	+	C	+	view	blast

Ensembl database

 [BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

 Search all species...

Human (GRCh37) ▾ Location: 3:12,392,625-12,393,625 Variation: rs1801282

Variation displays

Explore this variation

Genomic context

Genes and regulation (14)

Flanking sequence

Population genetics

Individual genotypes (2954)

Linkage disequilibrium

Phenotype Data (9)

Phylogenetic Context (6)

Citations (222)

External Data

LOVD

Configure this page

Add your data

Export data

Bookmark this page

Share this page

rs1801282 SNP

Original source

Alleles

Location

Co-located

Most severe consequence

Evidence status ⓘ

Clinical significance ⓘ

Synonyms ⊕

HGVS names ⊕

Genotyping chips ⊕

Explore this variation ⓘ

Genomic context

Genes and regulation

Population genetics

Individual genotypes

Linkage disequilibrium

Phenotype data

Citations

Individual genotypes

Phylogenetic context


Variants (including SNPs and indels) imported from dbSNP (release 138) | [View in dbSNP](#)


C/G | Ancestral: **C** | Ambiguity code: **S** | MAF: **0.07** (G)

Chromosome **3:12393125** (forward strand) | [View in location tab](#)

with **HGMD-PUBLIC CM981614**

Missense variant | [See all predicted consequences \(Genes and regulation\)](#)





This variation has **7** synonyms - click the plus to show

This variation has **17** HGVS names - click the plus to show

This variation has assays on **7** chips - click the plus to show

GENOTYPES

One copy of rs17822931 from the father and one copy from the mother

Copy 1: 5' ...GGCC**T**GAGT...3' (+)
 3' ...CCGG**A**CTCA...5' (-)

Copy 2: 5' ...GGCC**C**GAGT...3' (+)
 3' ...CCGG**G**CTCA...5' (-)

Genotype for rs17822931 on plus strand:

 T;C (T;C) C,T T,C (T,C)
 (C;T) CT C;T rs17822931 (T;C) TC
 rs17822931 (C;T)

Genotype for rs17822931 on minus strand:

 A;G (A;G) C,T A,G (A,G)
 (G;A) GA G;A rs17822931 (A;G) AG
 rs17822931 (G;A)

exercise

rs4788084

rs17822931

rs73546424

Copy 1:

5' . . . TCCC**C**TGGG . . . GGCC**T**GAGT . . . TGC**A**TGTGA . . . 3' (+)

3' . . . AGGG**G**ACCC . . . CCGG**A**CTCA . . . ACGT**A**CACT . . . 5' (-)

Copy 2:

5' . . . TCCC**C**TGGG . . . GGCC**C**GAGT . . . TGCA**A**GTGA . . . 3' (+)

3' . . . AGGG**G**ACCC . . . CCGG**G**CTCA . . . ACGT**T**CACT . . . 5' (-)

	<u>rs4788084</u> dbSNP orientation: minus	<u>rs17822931</u> dbSNP orientation: plus	<u>rs73546424</u> dbSNP orientation: plus
genotype on “plus strand”			
genotype on “minus strand”			
genotype on “dbSNP strand”			

HOMOZYGOTES AND HETEROZYGOTES

Homozygous:

Genotype consisting of two identical alleles at a given locus
(For a SNP: the same base at both copies, eg. C;C or A;A)

Heterozygous:

Genotype consisting of two different alleles at a locus
(For a SNP: the same base at both copies, eg. A;C)

exercise

rs4788084

rs17822931

rs73546424

Copy 1:

5' ... TCCC**C**TGGG ... GGC**C**TGAGT ... TGC**A**TGTGA ... 3' (+)

3' ... AGGG**G**ACCC ... CCG**G**A**C**TCA ... ACGT**A**CACT ... 5' (-)

Copy 2:

5' ... TCCC**C**TGGG ... GGCC**C**GAGT ... TGCA**A**GTGA ... 3' (+)

3' ... AGGG**G**ACCC ... CCG**G****G**CTCA ... ACGT**T**CACT ... 5' (-)



heterozygous
or
homozygous
?



heterozygous
or
homozygous
?

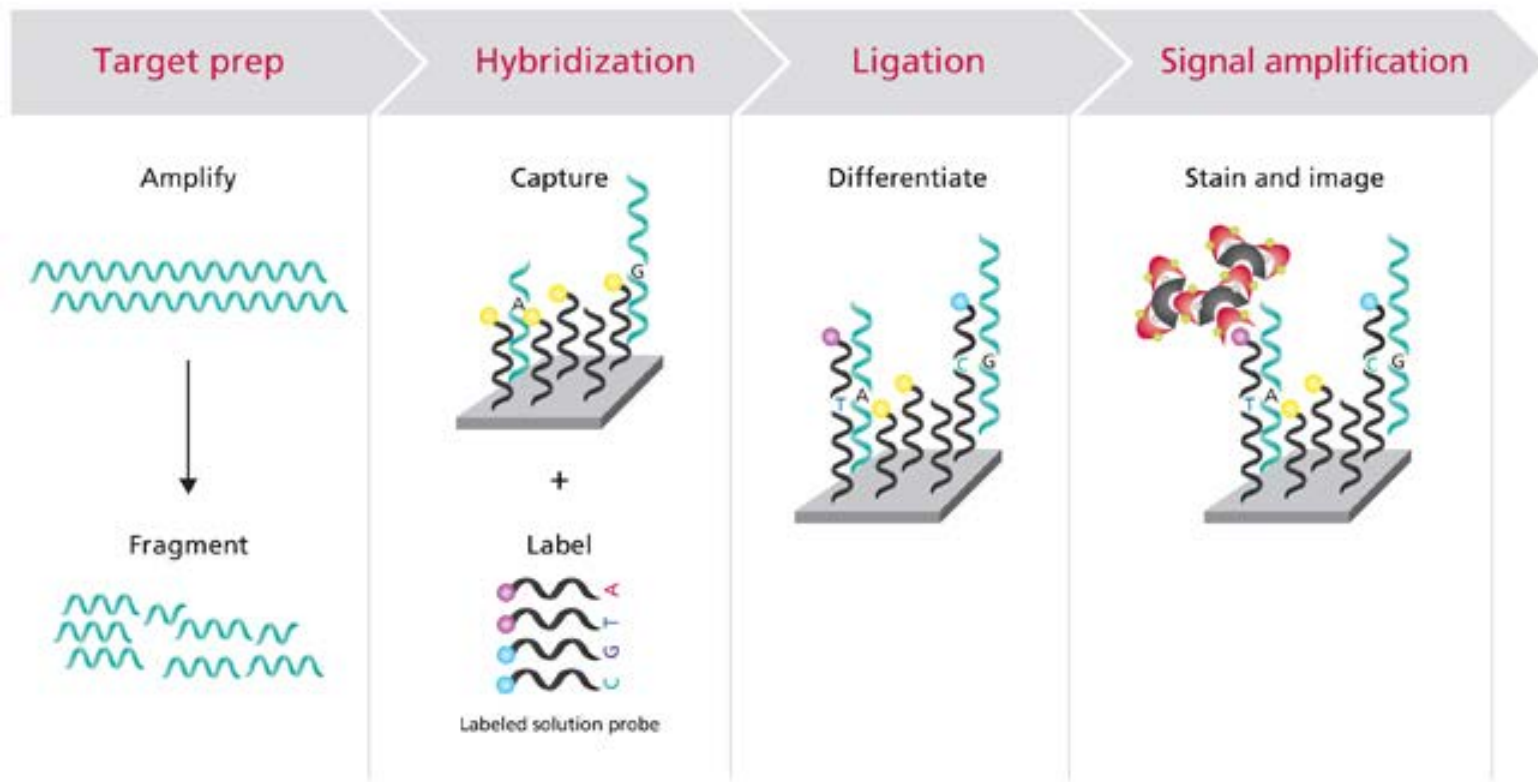


heterozygous
or
homozygous
?

GENOTYPING

SNP arrays

- Microarray technique
- 0.3 – 4.3 million SNPs



Genomic enlightenment Medicinsk Museion



Next Generation Sequencing

- Different parts of the genome can be sequenced:
 - ✧ Whole genome
 - ✧ Exome
 - ✧ Targeted
- Different methods for different platforms
- Is increasing in popularity due to increasingly lower costs



PHENOTYPES

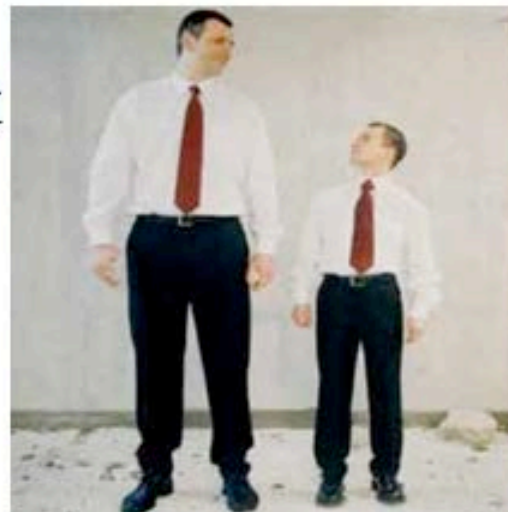
An observable characteristics or trait



eye color



height



Connecting genotypes and phenotypes:

GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Monogenic and polygenic traits

- Some traits are determined by a **single gene**, where a one mutation can cause a disease
 - Often called Mendelian diseases
 - Examples are Huntington's disease and sickle cell anemia
- Most common traits and diseases are caused by a **large number of genes**
 - Often called complex traits/diseases
 - Examples are human height, obesity, type 2 diabetes and cardiovascular disease
- **GWAS studies usually focus on complex polygenic traits**

GWAS

Association of common variants (SNPs) across the whole genome with a particular phenotype

Science 2007:

BREAKTHROUGH OF THE YEAR

Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

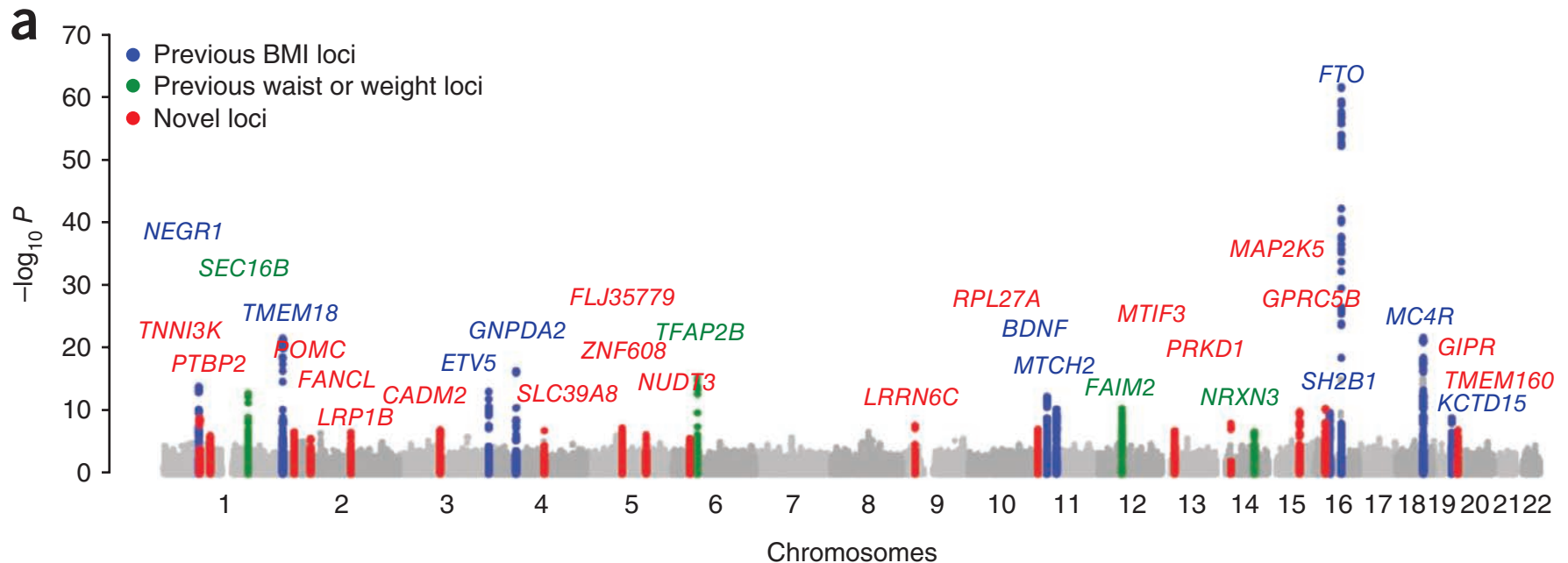


Cases vs controls

- Obtain DNA from a disease group (e.g. asthma) and a control group
- Obtain genotypes
- Identify variants that are significantly more common among cases than controls
- Those SNPs are associated with the disease (in this study)
- Not necessarily causal

Example of GWAS results (BMI)

Manhattan plot displays all SNPs on x-axis (ordered by chromosome)
and $-\log_{10}$ of p-values on y-axis

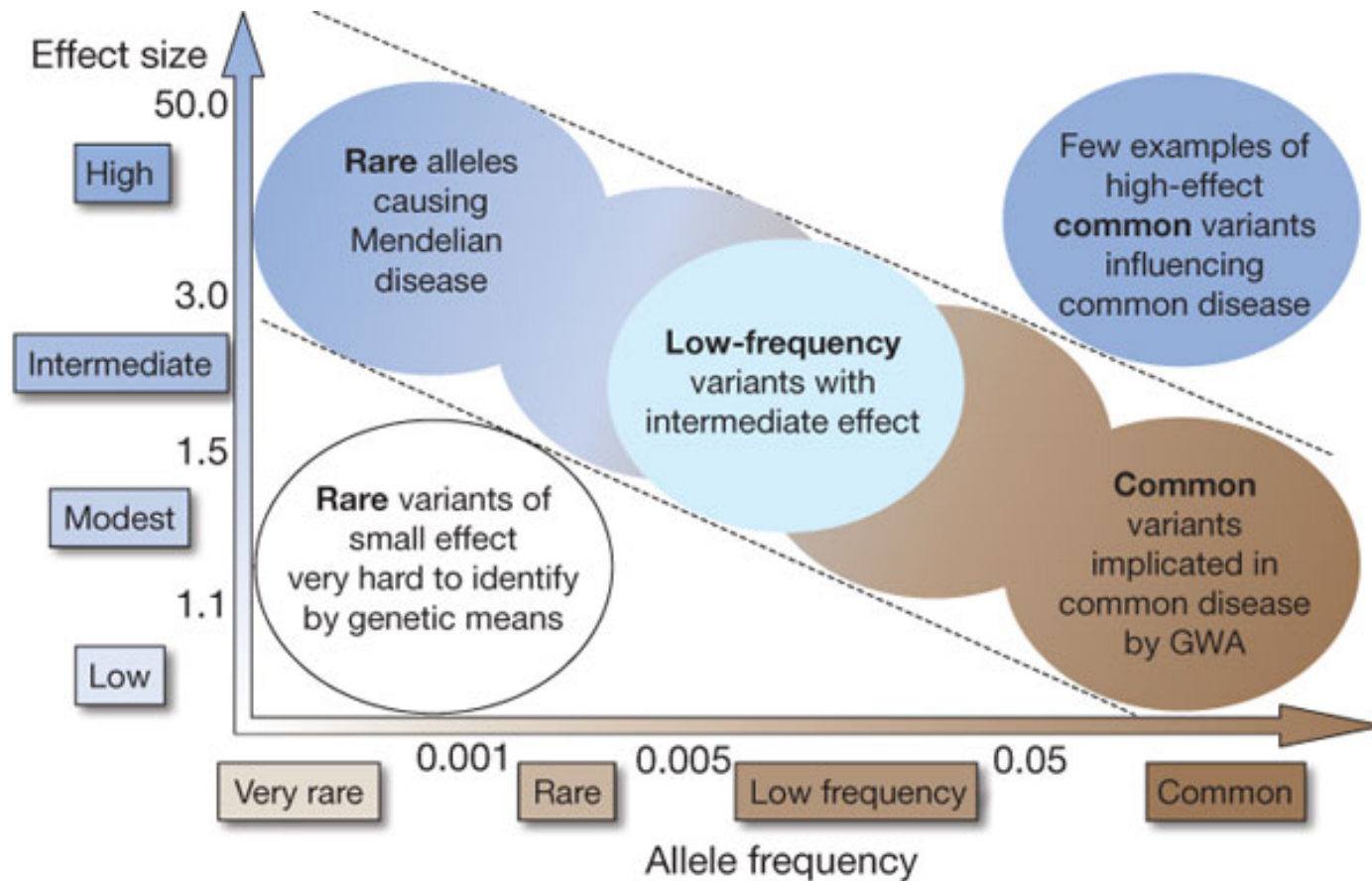


nature
genetics

Association analyses of 249,796 individuals reveal
18 new loci associated with body mass index

Speliotes et al, Nature Genetics, 2010

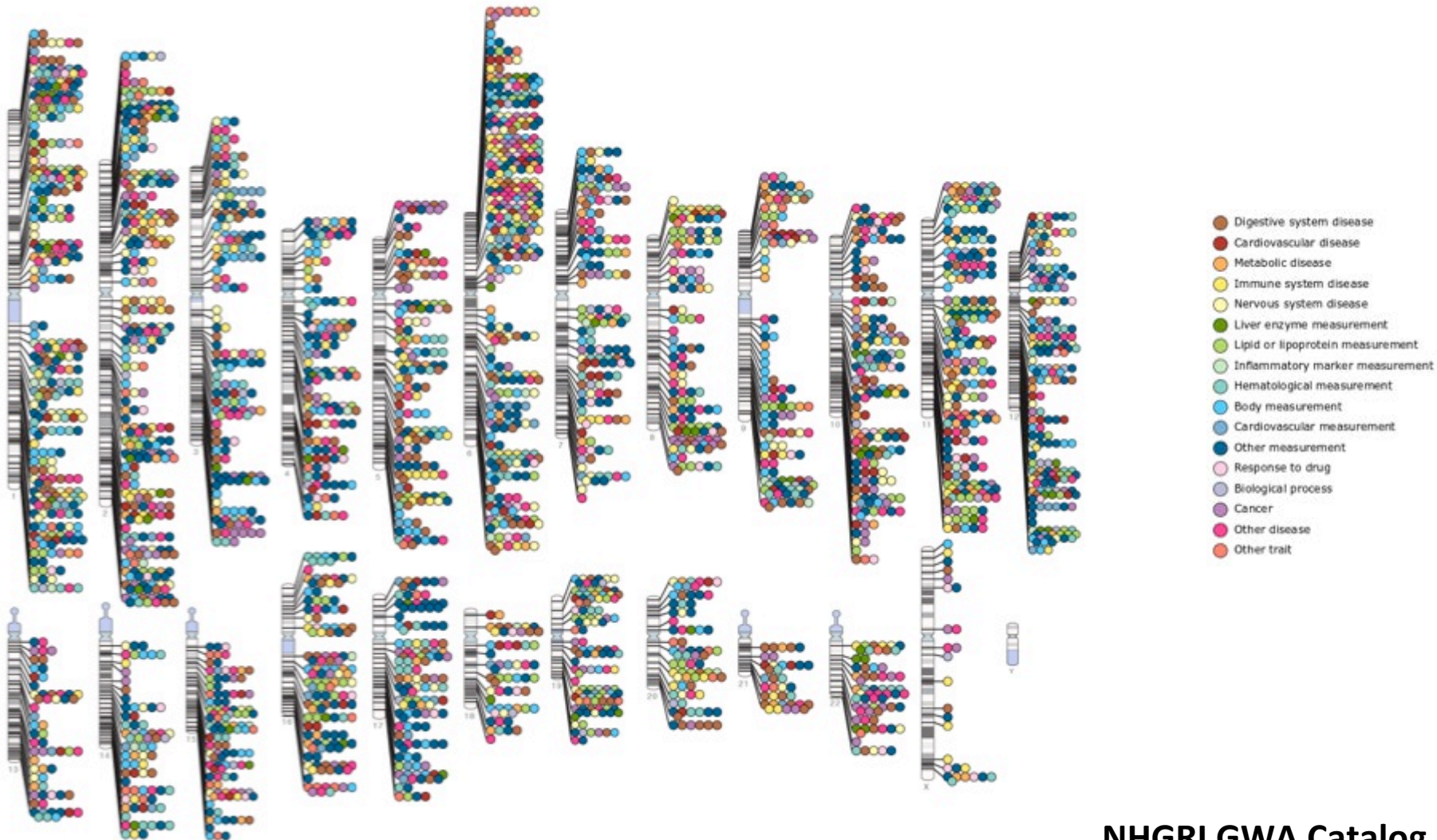
Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).



nature

NHGRI catalog

As of 07/01/14, the catalog includes 1782 publications and 12151 SNPs.



SNPedia.com

[Create account](#)  [Log in](#)

SNPedia


Page [Discussion](#)

[Read](#)

[Edit](#)

[View history](#)



[User:Lennon](#) answers questions for NPR's [On The Media](#) 

SNPedia

SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by [Promethease](#) to analyze and help explain your DNA.

Help! [\[edit\]](#)

- look at the example [rs1234](#)
- learn more about [SNPs](#)
- browse
 - [genes](#)
 - [genomes](#)
 - [genosets](#)
 - [genotypes](#)
 - [medicines](#)
 - [medical conditions](#)

[SNPedia](#)

[Promethease](#)

[FAQ](#)

[Blog](#)

[Recent changes](#)

[Random page](#)

▼ [Tools](#)

[What links here](#)

[Related changes](#)

[Special pages](#)

[Printable version](#)

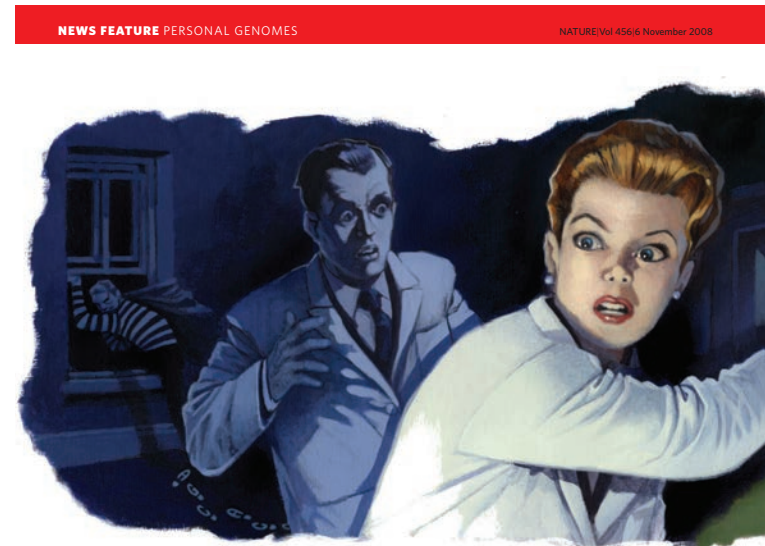
[Permanent link](#)

[Page information](#)

[Browse properties](#)

Challenges of GWAS

- Missing heritability
- Small effect sizes ($OR < 1.5$)
- Not much translation into clinical practice
- Biological role of variants unclear, majority (93%) outside of coding regions



The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

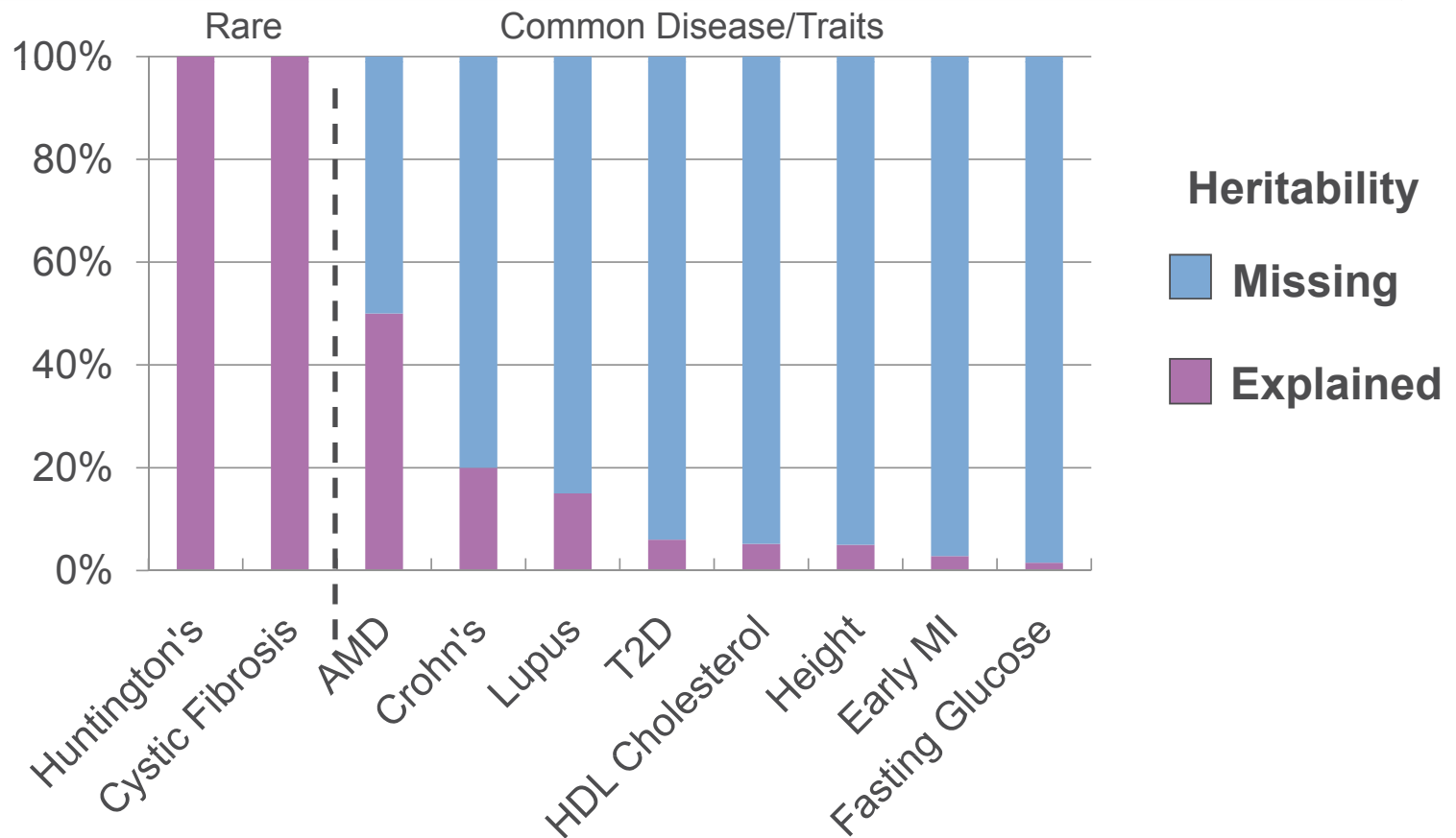
REVIEW

Five Years of GWAS Discovery

Peter M. Visscher,^{1,2,*} Matthew A. Brown,¹ Mark I. McCarthy,^{3,4} and Jian Yang⁵

The Case of the Missing Heritability

For most common diseases, the sum of individual effects found so far is much less than the total estimated heritability




GENOTYPE TO PHENOTYPE PREDICTIONS


It is difficult, perhaps impossible, to accurately predict complex traits from the information we have today!


- Clinical setting
 - High penetrance, often rare mutations, such as *BRCA1*
- Commercial genotyping
 - 23andMe
- Ancient genomes



Commercial genotyping

The largest DNA ancestry service in the world [sign in](#) [register kit](#)  **0**

 **23andMe** [welcome](#) [health](#) [ancestry](#) [how it works](#) [buy](#) [help](#)



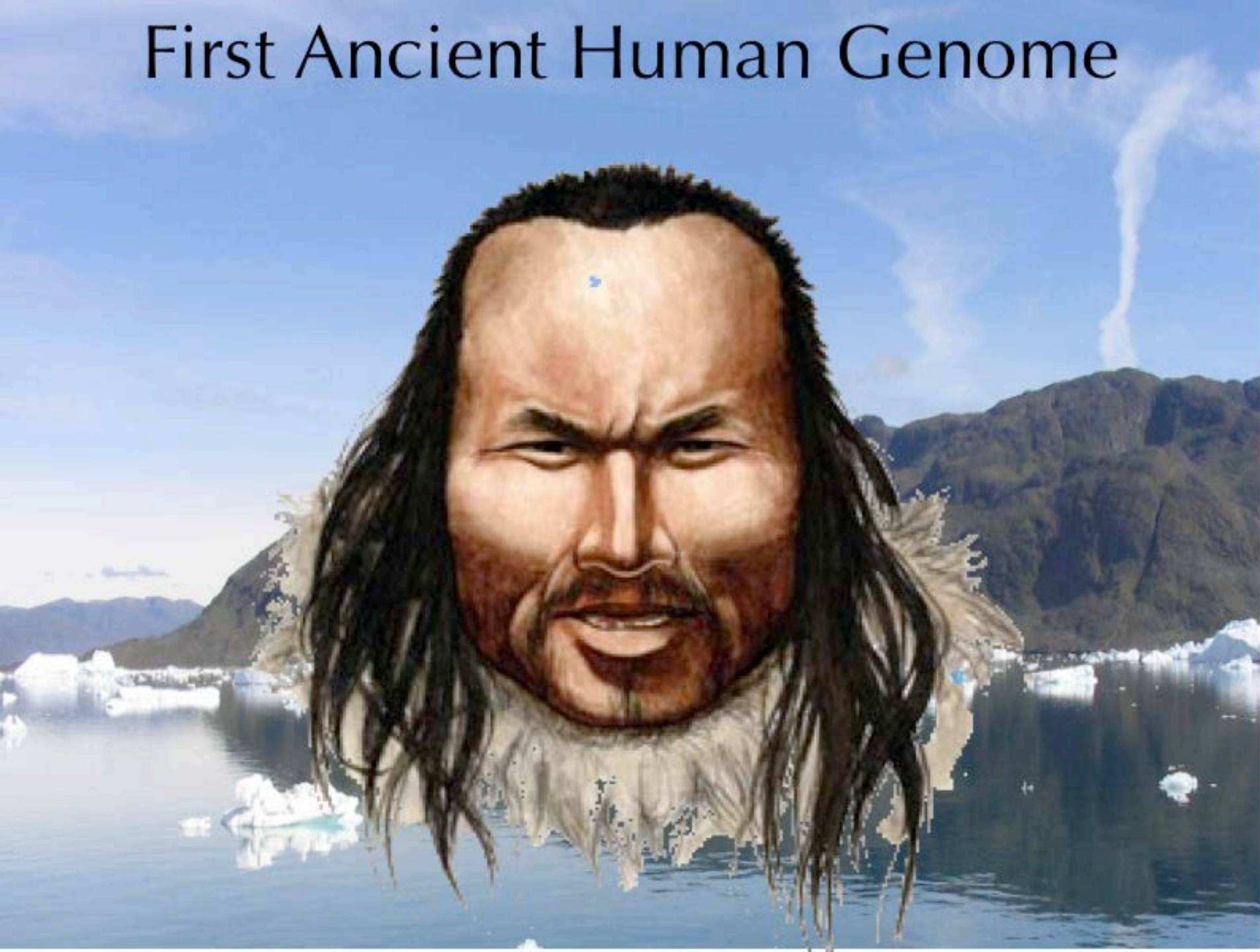
Discover your ancestral origins and lineage with a personalized analysis of your DNA.

- Learn what percent of your DNA is from populations around the world.
- Contact relatives across continents or across the street.
- Build your family tree and enhance your experience with relatives.

[order now](#) **\$99**

Genotype and phenotype
of an ancient genome

First Ancient Human Genome



The Saqqaq Genome Project

4,000 years

Hair sample from permafrost

DNA extraction <10% contamination

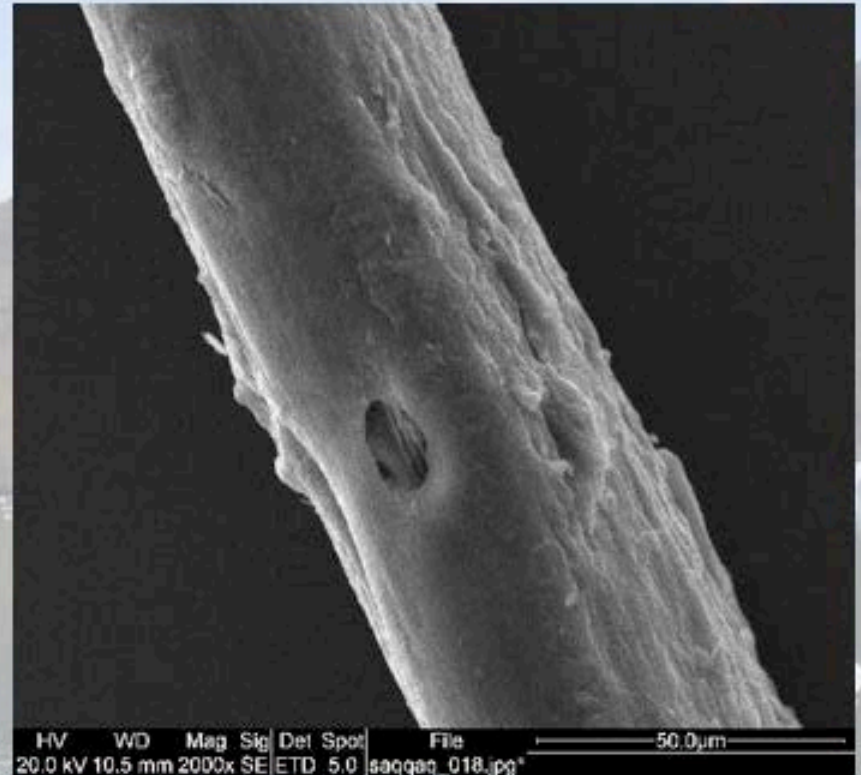
20 x coverage

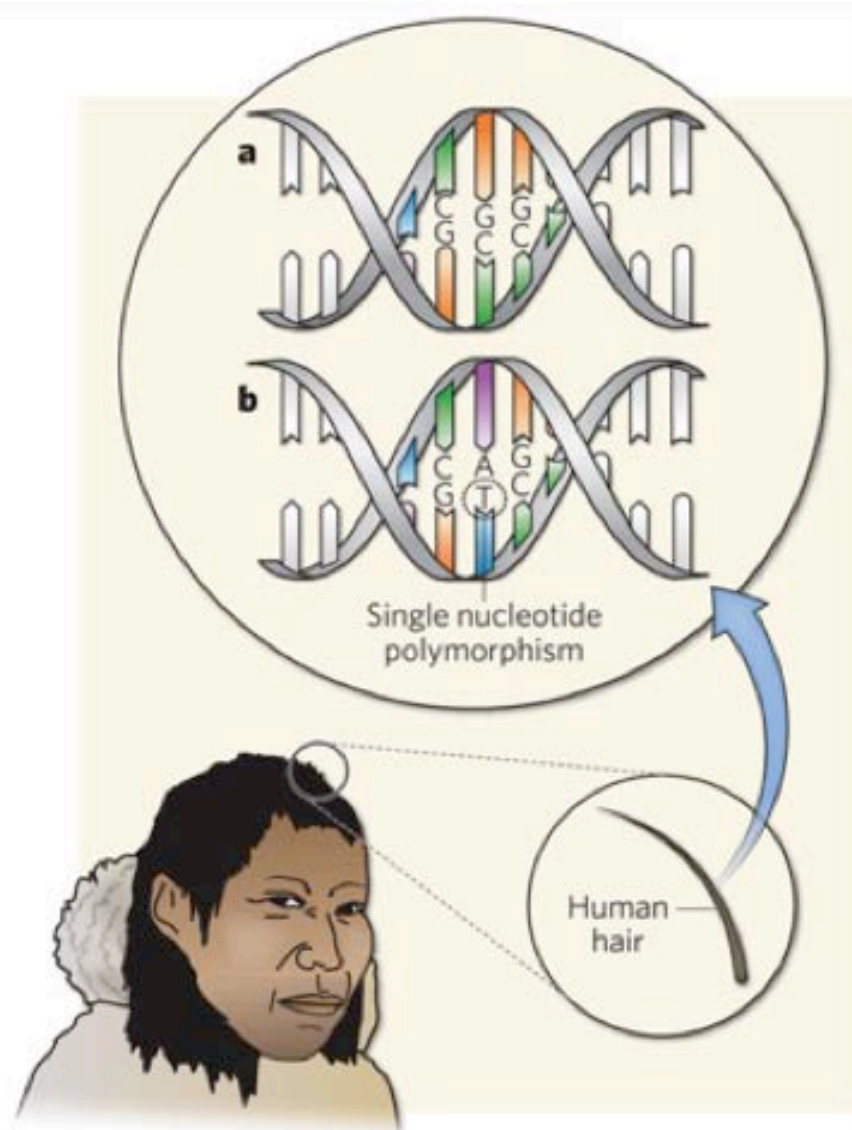
Started 2009



Eske Willerslev

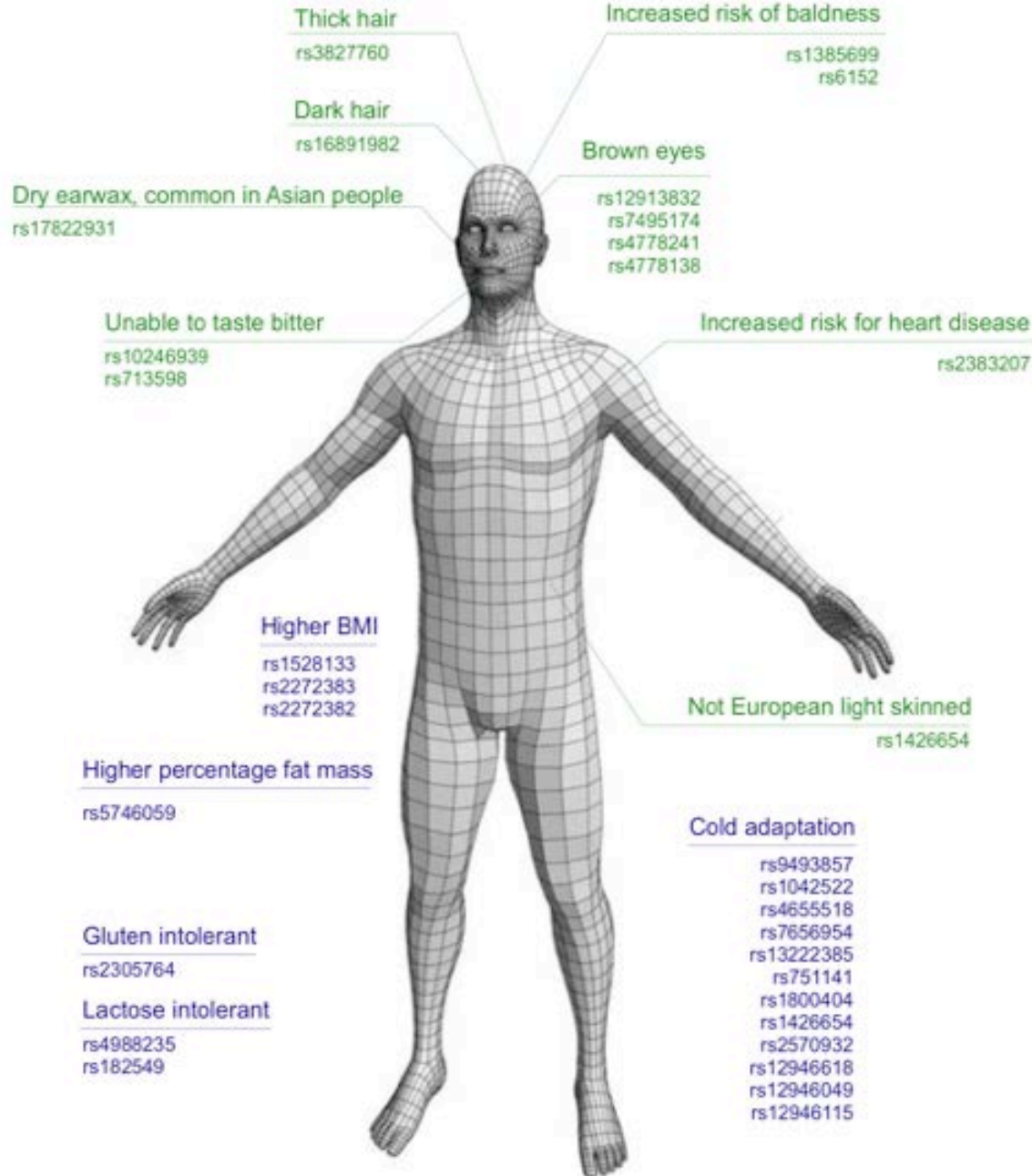
DNA from hair





Rasmussen *et al.*² have sequenced the genome of a man from the Saqqaa culture, using DNA from hair preserved in permafrost in Greenland. They analysed the genome to find single nucleotide polymorphisms (SNPs) — differences in single DNA base pairs that exist between individual genomes, and that may act as markers of an individual's physical traits. **a**, Here, a short stretch of human DNA is shown that is a marker for normal earwax. **b**, In the analogous DNA from the Saqqaa individual, there is a SNP in which a C in the lower strand has been replaced by a T (C, G, T and A denote the four kinds of DNA base). This SNP shows that the Saqqaa man had dry earwax. Rasmussen and colleagues identified other SNPs indicating that the ancient human had, among other things, brown eyes, non-white skin, thick dark hair and an increased susceptibility to baldness.

What can we say about his phenotypes?



From genotype to phenotype: how good are we at putting a face to an anonymous individual? - While some traits manifest themselves in a tissue specific manner (highlighted in green), others are more systemic (highlighted in blue). Going from the genetic blueprint to visual appearance, physiological behaviour and medical predispositions is still an open challenge.

The Saqqaq Genome Database (NCBI36)

Enter sequence range, identifier or cheat code

Examples

Range:	<code>17:398382..399882 (chromosome:start..end)</code>
SNP ID:	<code>rs17822931 or ENSSNP22423 - Ambiguously mapped SNPs and in-dels may return several records.</code>
List phenotypic associations on chromosome:	<code>1:phenotype</code>

Note: Query is currently limited to 100000 records/nucleotides

[[home](#) | [download flat files](#) | [ancientgenome.dk](#)]

The Saqqaq Genome Database (NCBI36)

Result

chr	pos	ref	is_ref	genotype	pp	depth	repeat	rs	type	strand	snp_alleles	trait	risk_allele	pmid	source
3	12368125	C	n	GG	0.0006439	11		rs1801282	single	+	CG	obesity, Obesity, association with, Disease-associated and putatively functional polymorphism, Type 2 diabetes	C	9792554 17463246 17463248 17463249	NHGRI, other

Columns explained

chr	The chromosome
pos	Position on chromosome
ref	The reference nucleotide on forward strand in hg18
is_ref	Indicates whether the genotype is the same as the reference nucleotide (y) or not (n)
genotype	The genotype called (on forward strand)
pp	For numerical reasons, we report (1-PP), where PP is the posterior probability of the genotype
depth	The number of reads covering the position
repeat	If the position lies in an annotated repeat, the ID is given here
rs	dbSNP rs-number
type	Type of dbSNP entry ("single", "indel" etc. - see the UCSC genome browser for details).
strand	Strand for dbSNP entry + (or 1) or - (or -1)
snp_alleles	Known SNP alleles, e.g. "AC" (or "A/C") for a SNP of type "single"
trait	Associated trait or phenotype (general information -- not dependant on this individual's genotype)
risk_allele	Allele associated with trait
pmid	Pubmed ID of GWAS paper reporting the association
source	Source database for association

EXERCISE

Genotype to phenotype exercise

[http://wiki.cbs.dtu.dk/teachingmaterials/
index.php/ExGenotype2PhenotypeLite](http://wiki.cbs.dtu.dk/teachingmaterials/index.php/ExGenotype2PhenotypeLite)